DOCUMENT RESUME

ED 377 246 TM 022 512

AUTHOR Crehan, Kevin D.; Haladyna, Thomas M.
TITLE A Comparison of Three Linear Polytomous Scoring Methods.
PUB DATE [94]
NOTE 25p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; Computer Assisted Testing; *Distractors (Tests); Models; *Multiple Choice Tests; *Pass Fail Grading; *Scores; Testing Programs; *Test Items
IDENTIFIERS *Linear Measurement; *Polytomous Scoring

ABSTRACT

More attention is currently being paid to the distractors of a multiple-choice test item (Thissen, Steinberg, and Fitzpatrick, 1989). A systematic relationship exists between the keyed response and distractors in multiple-choice items (Levine and Drasgow, 1983). New scoring methods have been introduced, computer programs developed, and research conducted to estimate and evaluate the potential for useful information present in distractors. This study determines the efficacy of three linear polytomous scoring methods in the context of testing programs where pass/fail decisions are made. Recommendations are offered about which methods appear to be most effective and feasible. Additionally, future directions in research on polytomous scoring are suggested. Four tables and one figure are included. (Contains 17 references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 377 246

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
KEVIN D. CREHAN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A Comparison of Three Linear Polytomous Scoring Methods[1]

Kevin D. Crehan
University of Nevada, Las Vegas
and
Thomas M. Haladyna
Arizona State University West

Address correspondence to:

Kevin D. Crehan
Counseling and Educational Psychology
College of Education
University of Nevada, Las Vegas
Las Vegas, NV 89154

TM022512

BEST COPY AVAILABLE

A Comparison of Three Linear Polytomous Scoring Methods

Abstract

More attention is currently being paid to the distractors of a multiple-choice test item (Thissen, Steinberg, & Fitzpatrick, 1989). A systematic relationship exists between the keyed response and distractors in multiple-choice items (Levine & Drasgow, 1983). New scoring methods have been introduced, computer programs developed, and research conducted to estimate and evaluate the potential for useful information present in distractors. This study examines the efficacy of three linear polytomous scoring methods in the context of testing programs where pass/fail decisions are made. Recommendations are offered about which methods appear to be most effective and feasible. Additionally, future directions in research on polytomous scoring are suggested.

Index terms or phrases:

multiple-choice items, polytomous scoring, licensure/certification exams, reliability, passing scores

Introduction

A systematic relationship exists between distractors and keyed answers in multiple-choice testing (Levine & Drasgow, 1983; Lord, 1977). This relationship has been the foundation for past attempts to use the differential information represented in wrong answer choice to score test results. The term "polytomous scoring" has been used to describe the use of this information.

Haladyna and Sympson (1988) have characterized polytomous scoring as consisting of two unique schools. The first is the more traditional and traces its roots back to the mid 1930s where Paul Horst was among the first to discuss the promise and potential of polytomous scoring. This class of methods involves the linear combination of option responses and option scoring weights. The resulting scales may be not necessarily linearly related to the original raw score scale, but clearly, the derivation of scores comes from the sum of weights over responses.

The second school consists of newer polytomous item response theories (Bock, 1972; Samejima, 1979; Sympson, 1981, 1983, 1986; Thissen & Steinberg, 1984). These methods are based on the use of option characteristic curves (trace lines) that define the scoring function through the range of ability or achievement being measured. To date, only one computer program, MULTILOG (Thissen, 1991), implements these theories and the program is limited to a small number of response categories. MULTILOG currently offers techniques for analyzing the effectiveness of small item sets rather than being a method for scoring tests. Research to date

using these models has focused on computerized adaptive testing, but, clearly, there is great potential in these non-linear methods for paper-pencil tests.

One limitation of these non-linear methods is the requirement of large samples to insure stable item parameter estimates. Since many testing programs involve less than 1,000 examinees, the use of these non-linear polytomous scoring methods is problematic. Another difficulty is the lack of software for implementing these theories. A third limitation, and one that applies to linear methods as well, is that the low quality of distractors for most test items limits the effectiveness of polytomous scoring (Haladyna & Downing, submitted for publication).

The present study examines the comparative efficacy and feasibility of three linear polytomous scoring methods with a data base from a large licensing examination. The study examines the relative merits of these methods with respect to score reliability, the consistency of pass/fail classification, and average absolute differences for ordered score groups between comparable test forms.

Approaches to Scoring Item Sets

In this section, the methods of scoring item sets are briefly discussed and evaluated.

Point-biserial

The relationship between dichotomous option score and total test score has been used for option weighting (Haladyna & Sympson, 1988). The point-biserial correlation has advantages of being accepted as an index of option performance and item analysis

programs routinely compute the point-biserial relationship between each option and total test score. Haladyna (1990) has shown that in a pass/fail setting, the point-biserial method works very effectively. Moreover, sample size requirements for stable estimates are much lower than for applications of item response theory.

Max-alpha

Guttman (1941) proposed a strategy to score multiple-choice test results using a theory that maximized coefficient alpha. This method uses the concept of "option mean", the mean of total test score for all examinees choosing an option. Each option's mean is used as an initial weight to score test results, a new total score, based on choice means, is used to recompute option means, and the process of iteration is continued to a criterion of stabilization in the change of coefficient alpha. Echternacht (1975) found that the initial option mean is very close to maximizing alpha and few iterations are needed. Therefore, the initial option mean is a relatively simple way to obtain an approximation. This option weighting strategy results in both differential option and item weighting with more difficult items having higher weights assigned to their keyed response.

One weakness of this option weighting procedure is the dependence of weights on the difficulty level of other items on the test. An item's keyed response would have a higher weight if presented with relatively easy items than if introduced within a more difficult item set. It is also possible for a keyed option to

have a lower weight than a distractor for the same item, although this outcome would likely lead to a decision to eliminate the item as would be the case for an item with a negative discrimination index.

Polyweighting

Sympson (1988) introduced both a method, polyweighting, and a computer program, POLY, that derives from the max-alpha technique but replaces option mean with percentile rank initial option weighting. As with the max-alpha, polyweighting results in items being weighted according to their difficulty, with correct responses to harder items given more weight than correct responses to easier items. The major advantage of using percentile ranks as option weights is the alleviation of sample dependence since the technique is analogous to equipercentile equating. Another advantage of polyweighting is that sample size requirements are lower than for applications of item response theory. Implementation of polyweighting using the computer program, POLY, is somewhat easy and the program has limits that are high enough to accommodate large testing programs.

Research on Linear Methods

The initial recorded research was done in the early 1940s by Guilford and his colleagues and this kind of research has been performed sporadically since. The general findings are that internal consistency is slightly improved with the use of these linear methods (Sympson & Haladyna, 1988), but correlations with external (validity) criteria do not improve. The explanation for

these findings is that the option weighting techniques tend to purify the trait measure through enhancing the internal consistency of the test results. So if the objective is to improve concurrent or predictive validity, these linear methods appear not to contribute, but if the objective is to make the trait more internally consistent, these linear methods are superior to traditional number-correct scoring.

Since polytomous scoring treats distractors individually rather than as a set, the quality and effectiveness of each distractor is an important consideration. A systematic relationship exists between the correct and incorrect answers (Levine & Drasgow, 1983; Thissen, Steinberg, & Fitzpatrick, 1989). Items that are effective in polytomous scoring must necessarily have distractors that work as intended. Despite the polytomous scoring method used, however, Haladyna and Downing (submitted for publication) found with four testing programs of differing design and use that few items yielded more than two effective distractors. If this work generalizes to most standardized tests, then the potential benefit of polytomous scoring will be limited until better distractors are developed.

Method

Four scoring methods are used in this study: (1) dichotomous one-zero scoring, (2) the point-biserial option weight, (3) the simple option mean variation of max-alpha, and (4) Sympson's polyweighting.

The data are responses of 3,000 examinees administered the

National Association of Boards of Pharmacy Licensing Examination (NABPLEX). This test is constructed from test specifications using items that have been previously field tested. High standards for test development, item writing, and test assembly are maintained.

This test consists of two 150-item multiple-choice sections, correspond to morning and afternoon testing sessions that will subsequently be called Split 1 and Split 2. The sample was divided into two randomly equal halves each consisting of 1,500 examinees that will subsequently be called Sample A and Sample B. A double-cross validation design was used to assess the stability of option weights. For each polytomous scoring technique total score for each 150-item form was computed using the weights from the other sample. For the conventional number-correct, similar cross validation is not informative since the option weights of one and zero are the same in both samples.

Correlations within each test split for both samples were determined among scoring methods and cross-validations. Correlations between test splits for each scoring method were also computed as indices of a parallel forms type of reliability estimate.

Internal consistently (alpha) reliability estimates were computed for each scoring method and cross-validation.

To assess consistency of pass-fail decisions among the scoring methods, simulated passing scores were established in seven locations in the test score distributions using the proportional cuts of 1/5, 1/3, 2/5, 1/2, 3/5, 2/3, and 4/5, i.e., the cut at 1/5

"passes" 80% of the examinees. Each examinee was scored pass-fail for the seven cut points separately for test splits 1 and 2. A percentage of consistent decisions was then determined for each scoring method and cross-validation. This was done to generate information about the comarative suitability of these four scoring methods in the various potential passing scores ranges used in testing programs where pass/fail decisions are made.

To further assess the relative precision among the scoring methods, average absolute differences were calculated between test splits by quintile. Standard z-scores were calculated for each sample and test split. Cases for test split 1 were grouped into five ordered score categories by dividing the distribution at the four quintiles. Absolute z-score differences were determined between split 1 and split 2 for each case. These absolute differences were then averaged within each of the five ordered score categories.

Descriptive indices of skewness and kurtosis were also determined to assess the change in these distributional characteristics between the raw-score and option weighted score distributions.

Results and Discussion

Correlations among the scores derived from these scoring methods were quite high with more than half the coefficients over .98. There was no consistent pattern in the correlation matrices over samples and test splits. Cross-validation correlations were extremely high with eight of the twelve coefficients being above

.995. All four of the polyweighting correlations exceeded .995.

****

insert Table 1 here

****

Comparable forms reliabilities are reported in Table 1. The only consistent observation slightly favors polyweighting in both the original scoring and cross-validations.

****

insert Table 2 here

****

Table 2 presents alpha reliabilities for each scoring method by test split and sample. Raw score reliabilities aggregate to about .87 and the three polyweighting methods aggregate to about .90. The difference in reliability of .03 equates to lengthening a test with .87 reliability by about 35% to attain a reliability of .90 or, conversely, a reduction in length of 26% for a test with .90 reliability to maintain a reliability of .87. This finding is consistent with previous evidence concerning the comparable reliability of linear option-weighted and number-correct scores. It is notable that the reliabilities of the cross-validation samples are consistently higher than the raw-score reliabilities.

****

insert Table 3 here

****

Results for the percent consistent decisions analysis are presented in Table 3. On average, choice mean and polyweighted

scores had about 1.3% higher consistency in classification than raw-scores and outperformed the point-biserial method by an even greater margin. Further inspection of Table 3 reveals a higher overall consistency at the extreme 1/5 and 4/5 cuts with no apparent differences between the raw-score and choice mean or polyweighted scores but with the point-biserial score demonstrating about a 1.4% lower consistency than the other scoring methods. Aggregating results for the middle five passing proportional cuts (1/3 to 2/3 inclusive) shows an observed difference of about 2% consistency between raw-scores and choice mean or polyweighted scores and no difference between raw-scores and point-biserial scores.

The general pattern of consistency appears to maintain in cross-validation albeit with the rate of consistency for point-biserial cross-validation faring better than the original point-biserial scores.

****

insert Table 4 here

****

Table 4 and Figure 1 summarize and portray the results of the average absolute difference analysis. Figure 1 illustrates an overall decrease in average absolute differences from the first to fifth ordered score groups. The raw-score and point-biserial score differences decrease at about the same rate while the option mean and polyweight score differences decrease at a higher rate over the

score groups. There is no discernable difference among scoring methods for examinees in the first score group. However, a comparison between the combination of raw-score and point-biserial with the combination of option mean and polyweights for the fifth score group shows a difference of about .09, that is about one-third of the raw-score standard deviation for this score group.

Finally, the converting from raw-scores to option mean and polyweighted scores had the effect of increasing the negative skew of the original raw score distribution and, also, increasing kurtosis, i.e., moving toward a more leptokurtic shape. This effect was not present for the point-biserial conversion.

Conclusions

As has been observed in previous research, option weighting has the effect of increasing the internal consistency of the score distribution. In the present study this increase was equivalent to an increase in test length of about one-third. Since there was no external validity criterion against which to compare scoring methods in this study, the question of effects of option weighting on criterion related validity are not addressed. However, if one assumes that the test items used in this study are a good representation a larger domain of possible items, then an argument for enhanced content coverage can be made.

Of even greater interest to the practitioner are the findings related to consistency in selection classifications. On average, choice mean and polyweighting had 1.3% higher classification consistency than raw-scores and the rate of consistency was even

higher (2%) for the cut score proportions between one-third and two-thirds inclusive. Although these differences are small, they translate to 39 to 60 fewer consistent classifications based on raw-scores for a sample as large as the one used in this study.

Perhaps the most interesting observation in this study is the results of the average absolute difference analysis. For the item data used in this analysis, the differential between raw-scores and choice mean or polyweight scores in average absolute difference increased over the ordered score groups to one-third standard deviation difference for the highest score group. This suggests an increase in score precision with polytomous scoring, using either choice mean or polyweighting, as test scores increase. This result is, of course, likely to differ with other test score distributions. The test results used in this study had items with an average of 75% correct.

Applicability to Testing Programs

Should a testing program designer risk a venture into polytomous scoring? The nearly 50 years of research on this topic may suggest that if internal consistency of the trait measure is a primary concern, then almost any polytomous scoring method will outperform a dichotomous scoring method.

Many licensing and certification testing programs depend primarily on a test meeting certain content specifications that are originally developed from a job analysis, task analysis, role delineation study, or practice analysis. In these contexts, sampling of content is critical to test score interpretation. In

such settings, the polytomous scoring techniques may work, but problems exist if the content domain is not homogeneous.

If the test is used for selection or placement, such as with a college or graduate admissions test, the polytomous scoring may be inappropriate because it tends to reduce predictive and concurrent correlations, due to the characteristic of making the trait more internally consistent.

Finally, as has been pointed out earlier in this paper, the quality of distractors in most tests does not seem to be high enough to benefit fully from polytomous scoring. If the findings of Haladyna and Downing (submitted for publication) can be generalized, it would appear that few items exist with more than two distractors worthy of polytomous scoring methods.

References

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37, 29-51.

Echternacht, G. (1975). The variances of empirically derived option scoring weights. Educational and Psychological Measurement, 36, 301-309.

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In P. Horst (Ed.) Prediction of Personal Adjustment. Social Science Research Bulletin 48, 321-345.

Haladyna, T. M. (1990). Effects of empirical option weighting on estimating domain scores and making pass/fail decisions. Applied Measurement in Education, 3, 231-244.

Haladyna, T. M., & Downing, S. M. (submitted for publication). How many options is enough for a multiple-choice test item?

Haladyna, T. M., & Sympson, J. B. (1988). Empirically based polychotomous scoring of multiple-choice test items: A review. Paper presented in C. E. Davis (Chair), New Developments in Polychotomous Item Scoring and Modeling. Symposium conducted at the annual meeting of the American Educational Research Association, New Orleans, LA.

Levine, M. V., & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. Educational and Psychological Measurement, 43, 675-685.

Lord, F. M. (1977) Optimal number of choices per item: A comparison of four approaches. Journal of Educational Measurement, 14, 33-38.

Samejima, F. (1979). A new family of models for the multiple-choice item. Office of Naval Research Report 79-4. Knoxville, TN: University of Tennessee.

Sympson, J. B. (1981). A nominal model for IRT item calibration. Talk given at the Office of Naval Research Conference on Model-based Psychological Measurement, Millington, TN.

Sympson, J. B. (1983). A new item response theory model for calibrating multiple-choice items. Paper presented at the annual meeting of the Psychometric Society, Los Angeles, CA.

Sympson, J. B. (1986). Extracting information from wrong answers in computerized adaptive testing. Paper presented in C. E. Davis (Chair), New Developments in Polychotomous Scoring. Symposium

conducted at the annual meeting of the American Educational Research Association, Chicago, IL.

Sympson, J. B. (1988). A procedure for linear polychotomous scoring of test items. Paper presented at the Office of Naval Research Contractors' Meeting on Model-based Psychological Measurements, Iowa City, Iowa.

Sympson, J. B., & Haladyna, T. M. (1988). An evaluation of "polyweighting" in domain-referenced testing. Paper presented in C. E. Davis (Chair), New Developments in Polychotomous Item Scoring and Modeling. Symposium conducted at the annual meeting of the American Educational Research Association, New Orleans, LA.

Thissen, D. (1991). Multilog. Chicago, Il.: Scientific Software, Inc.

Thissen, D. & Steinberg, L. (1984). A response model for multiple-choice items. Psychometrika, 49, 501-519.

Thissen, D., Steinberg, L. & Fitzpatrick, A. R. (1989). Multiple-choice Models: The distractors are also part of the item. Journal of Educational Measurement, 26, 161-176.

Table 1

Comparable Forms Reliabilities of Each Scoring Method and Cross Validations for Two Samples (N = 1500 each)

Scoring Method	Sample A	Sample B
Raw-Score	.8551	.8345
Point-Biserial	.8717	.8107
Option Mean	.8545	.8487
Polyweights	.8768	.8615
<u>Cross-Validations</u>		
Point-Biserial	.8709	.8557
Option Mean	.8301	.8370
Polyweights	.8768	.8620

Table 2
Coefficient Alpha Reliabilities for Both Subject Samples and Both Test Splits

Scoring Method	Test Split	Sample A	Sample B
Raw-score	1	.879	.868
	2	.877	.872
Point-Biserial	1	.892	.883
	2	.898	.891
Choice Mean	1	.916	.902
	2	.914	.908
Polyweights	1	.907	.897
	2	.909	.903
<u>Cross-Validations</u>			
Point-Biserial	1	.886	.878
	2	.889	.887
Choice Mean	1	.888	.882
	2	.886	.881
Polyweights	1	.897	.889
	2	.901	.895

Table 3

Percent consistant decisions for two test splits based on proportional cuts of one-fifth, one-third, two-fifths, one-half, three-fifths, two-thirds, and four-fifths of examinees averaged over two samples (N = 1500 each).

	Proportional Cuts							
Scoring Methods	1/5	1/3	2/5	1/2	3/5	2/3	4/5	Avg
Raw-score	89.2	83.4	82.8	81.7	83.1	83.7	88.2	84.6
Point-Biserial	88.0	83.6	82.7	82.1	81.6	84.2	86.8	84.1
Choice Mean	89.5	85.6	84.6	83.9	84.6	85.6	88.5	86.0
Polyweights	88.8	85.3	84.8	83.2	85.0	86.0	88.4	85.9
Cross-Validations								
Point-Biserial	89.3	84.4	83.2	83.4	83.7	84.6	88.4	85.3
Choice Mean	88.9	85.2	83.8	83.3	84.0	84.9	88.2	85.5
Polyweights	89.2	85.5	84.8	83.6	84.2	85.2	88.3	85.8

Table 4

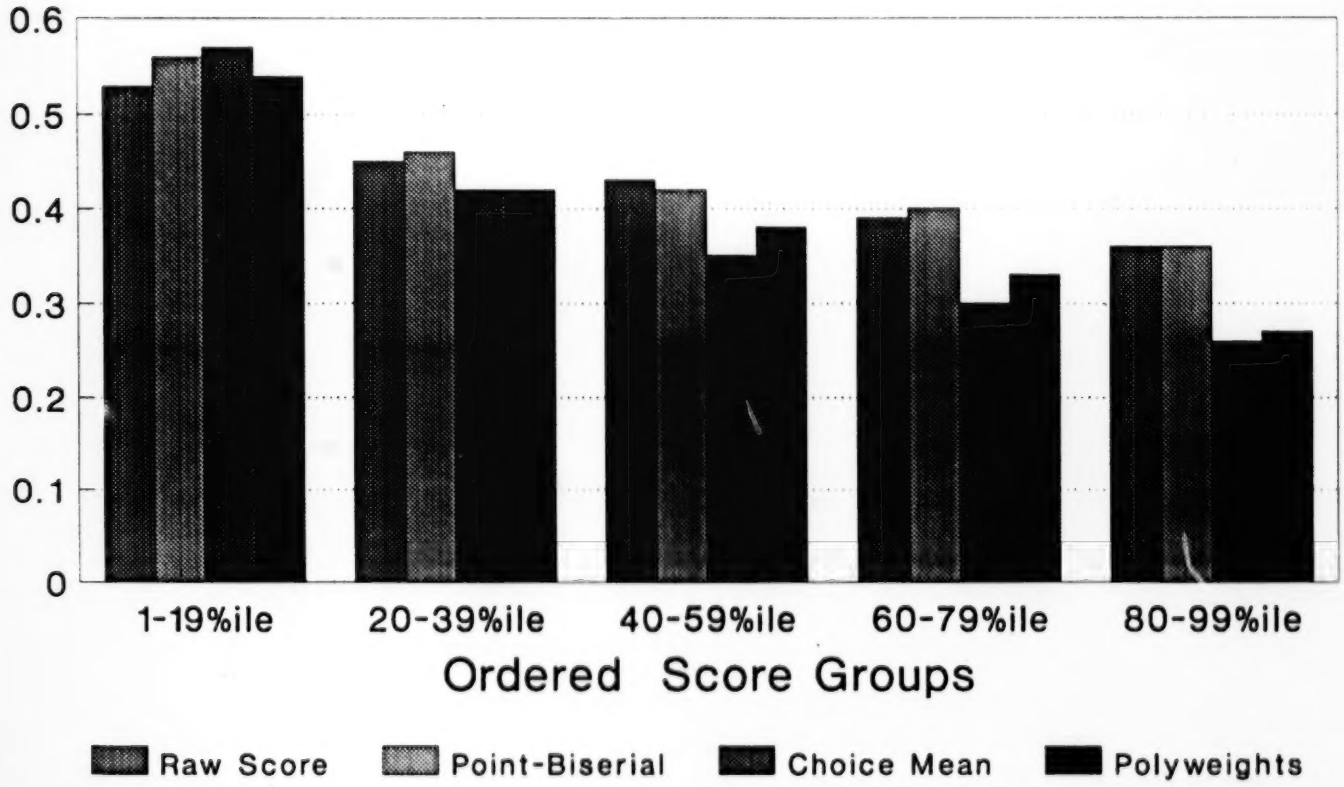
Average Absolute Differences Between Two Test Splits for Five Ordered Score Groups Reported as the Mean (SD) of Two Samples (N = 1500 each)

Scoring Method	Ordered Score Groups First	Second	Third	Fourth	Fifth	Avg
Raw Score	.53 (.42)	.45 (.34)	.43 (.39)	.39 (.30)	.36 (.28)	.43 (.35)
Point-Biserial	.56 (.35)	.46 (.35)	.42 (.32)	.40 (.29)	.36 (.27)	.44 (.35)
Choice Mean	.57 (.57)	.42 (.41)	.35 (.27)	.30 (.25)	.26 (.21)	.38 (.39)
Polyweights	.54 (.42)	.42 (.36)	.38 (.28)	.33 (.26)	.27 (.22)	.39 (.33)
Cross-Validations						
Point-Biserial	.51 (.39)	.42 (.33)	.41 (.34)	.36 (.27)	.32 (.25)	.41 (.33)
Choice Mean	.58 (.61)	.45 (.44)	.36 (.29)	.32 (.27)	.27 (.24)	.40 (.42)
Polyweights	.54 (.42)	.42 (.36)	.37 (.28)	.33 (.25)	.29(.24)	.39(.33)

Figure 1

Average Absolute Difference for Five Ordered Score Groups

Average Absolute Difference
For Five Ordered Score Groups



[1]Completion of this research was greatly facilitated by a sabbatical leave granted to the first author by the University of Nevada, Las Vegas

Address correspondence to:

Kevin D. Crehan
Counseling and Educational Psychology
College of Education
University of Nevada, Las Vegas
Las Vegas, NV 89154